Cheap, fast, and good enough for the non-biomedical domain but is it usable for clinical Natural Language Processing? Evaluating crowdsourcing for clinical trial announcement named entity annotations.

Haijun Zhai*, PhD, Todd Lingren*, MA, Louise Deleger, PhD, Qi Li, PhD, Megan Kaiser, BA, Laura Stoutenborough, BSN, Imre Solti‡, MD, PhD, MA

Division of Biomedical Informatics,Cincinnati Children's Hospital Medical Center Cincinnati, OH, USA

* Equal contribution ‡ Senior and corresponding author

Haijun.Zhai, Todd.Lingren, Louise.Deleger, Qi.Li, Megan.Kaiser, Laura.Stoutenborough, Imre.Solti@cchmc.org

***Abstract*— Building upon previous work from the general crowdsourcing research, this study investigates the usability of crowdsourcing in the clinical NLP domain for annotating medical named entities and entity linkages in a clinical trial announcement (CTA) corpus. The results indicate that crowdsourcing is a feasible, inexpensive, fast, and practical approach to annotate clinical text (without PHI) on large scale for medical named entities. The crowdsourcing program code was released publicly.**

I. Introduction

To reduce the cost of expert human annotation, many projects in general NLP have turned to crowdsourcing, which involves submitting a large number of smaller subtasks to a coordinated marketplace of workers on the internet. Snow et al. [1] were the first to explore the feasibility of crowdsourcing in NLP. Only a handful studies investigated crowdsourcing in the biomedical domain, usually on very small pilot sample sizes. The goal of our study was to evaluate the usability of crowdsourcing approach in the clinical NLP domain. The clinical NLP tasks that we used for the purpose of evaluation were medical named entity recognition and entity linking in a clinical trial announcement (CTA) corpus. First, we studied the turkers' performance to annotate medical named entities on a large scale. Second, we proposed to use crowdsourcing for named entity linking (e.g. to link <medication name, attribute> pairs). Third, we implemented a novel solution to improve the quality of manually created gold standard by an iterative model of crowdsourced error-correction.

II. Data and Methods

To build the gold standard for evaluating the crowdsourcing workers' performance, 1042 CTAs from the ClinicalTrials.gov website were randomly selected and double annotated for medication names, medication types and linked attributes. CrowdFlower, an Amazon Mechanical Turk-based crowdsourcing platform was utilized for the project. A named entity annotation friendly Graphical User Interface was programmed in JavaScript to leverage CrowdFlower's Markup Language (CML). We calculated F-measure to evaluate the quality of the annotated corpus and tested the statistical significance ($p<0.001$ Chi-square test).

III. Results

Medical named entities annotation task: The IAA F-measure of medication name and medication type were 0.871 and 0.729, respectively.

Correction task: The IAA F-measure of medication name and medication type achieved 0.900 and 0.760, respectively. With comparison to the F-measure of its corresponding correction baseline, improvements of 2.62 percent (0.877) and 10.79 percent (0.686) were gained.

Linking task: Non-expert annotators (turkers) did a very good job, in which the F-measure achieved 0.970.

IV. Discussion

Strict quality control (based on in-house generated small batch of gold-standard data), user friendly annotation interface, and continuous monitoring of annotator performance are critical for a successful clinical NLP crowdsourcing project.

V. Conclusions

Crowdsourcing is a feasible, inexpensive, fast, and practical approach to annotate clinical text without PHI on large scale for medical named entities. We publicly released the JavaScript and CML infrastructure codes that are necessary to utilize CrowdFlower's quality control and crowdsourcing interfaces for named entity annotations (https://code.google.com/p/soltilab/).

Acknowledgment

This work was supported by internal funds from Cincinnati Children's Hospital Medical Center. IS and LD were partially supported by grant 5R00LM010227-04.

References

[1] R. Snow , B. O'Connor , Daniel Jurafsky , Andrew Y. Ng, Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254-263, 2008.

978-0-7695-4921-7/12 $26.00 © 2012 IEEE
DOI 10.1109/HISB.2012.31